Chapter 3

Hypothesis testing

3.1 Statistical hypothesis testing

In Chapter 2, we discussed one component of statistical inference, estimating population parameters. We also introduced the philosophical and statistical differences between frequentist and Bayesian approaches to parameter estimation. The other main component of inference, and one that has dominated the application of statistics in the biological sciences, is testing hypotheses about those parameters. Much of the philosophical justification for the continued use of statistical tests of hypotheses seems to be based on Popper's proposals for falsificationist tests of hypotheses (Chapter 1). Although Jerzy Neyman, Egon Pearson and Sir Ronald Fisher had developed their approaches to statistical testing by the 1930s, it is interesting to note that Popper did not formally consider statistical tests as a mechanism for falsifying hypotheses (Mayo 1996). Hilborn & Mangel (1997, pp. 15–16) stated that "Popper supplied the philosophy and Fisher, Pearson, and colleagues supplied the statistics" but the link between Popperian falsificationism and statistical tests of hypotheses is still controversial, e.g. the contrasting views of Mayo (1996) and Oakes (1986). We will present a critique of statistical hypothesis tests, and significance tests in particular, in Section 3.6.

The remainder of this section will provide an overview of statistical tests of hypotheses.

3.1.1 Classical statistical hypothesis testing

Classical statistical hypothesis testing rests on two basic concepts. First, we must state a statistical null hypothesis (H_0), which is usually (though not necessarily) an hypothesis of no difference or no relationship between population parameters (e.g. no difference between two population means). In many cases, we use the term effect to describe a difference between groups or experimental treatments (or a non-zero regression slope, etc.), so the H_0 is usually an hypothesis of no effect. The philosophical basis for the statistical null hypothesis, at least in part, relates back to Popperian falsificationism, whereby science makes progress by severely testing and falsifying hypotheses. The implication is that rejection of the statistical H_0 is equivalent to falsifying it and therefore provides support ("corroboration") for the research hypothesis as the only alternative (Underwood 1997). We do not test the research hypothesis in this way because it is rarely more exact than postulating an effect, sometimes in a particular direction. Fisher (1935) pointed out that the null hypothesis is exact, e.g. a difference of zero, and is the result we would expect from randomizing observations to different experimental groups when there is no effect of the experimental treatment (Mulaik *et al.* 1997). The philosophical justification for testing the null hypothesis is still a controversial issue. For example, Oakes (1986) argued that support for the research hypothesis as a result of the null being rejected is not true corroboration and statistical tests, as currently practiced, have only superficial philosophical respectability.

Second, we must choose a test statistic to test the H_0. A test statistic is a random variable and, as such, can be described by a probability distribution. For example, a commonly used test statistic

for testing hypotheses about population means is t, where:

$$t = \frac{(\bar{y} - \mu)}{s_{\bar{y}}} \quad (3.1)$$

We introduced the t statistic and its probability distribution in Chapters 1 and used it in Chapter 2 for determining confidence intervals for population means. Test statistics like t have a number of probability distributions (see Figure 1.2), called sampling distributions, one for each possible degrees of freedom ($n - 1$). These sampling distributions represent the probability distributions of t based on repeated random sampling from populations when the H_0 is true and are sometimes called central distributions. Probabilities associated with particular ranges of values of test statistics are tabled in most statistics textbooks. Note that test statistics are continuous random variables, so we cannot define the probability of a single t value, for example. We can only talk about the probability that t is greater (or less than) a certain value or that t falls in the range between two values.

Before we look at the practical application of statistical tests, some consideration of history is warranted. The early development of statistical hypothesis testing was led primarily by Sir Ronald Fisher, whose influence on statistics was enormous. Fisher (1954, 1956) gave us null hypothesis or significance testing in statistics with the following methodology (Huberty 1993).

1. Construct a null hypothesis (H_0).

2. Choose a test statistic that measures deviation from the H_0 and that has a known sampling distribution (e.g. t statistic).

3. Collect the data by one or more random samples from the population(s) and compare the value of the test statistic from your sample(s) to its sampling distribution.

4. Determine P value, the associated probability of obtaining our sample value of the statistic, or one more extreme, if H_0 is true

5. Reject H_0 if P is small; retain H_0 otherwise.

Fisher proposed that we should report the actual P value (e.g. $P = 0.042$), which is a property of the data and could be viewed as a "strength of evidence" measure against H_0 (Huberty 1994). Fisher also introduced the idea of a conventional probability (of obtaining our sample data or data more extreme if H_0 is true) for rejecting H_0; this is called a significance level. He suggested a probability of one in twenty (0.05 or 5%) as a convenient level and the publication of tables of sampling distributions for various statistics reinforced this by only including tail probabilities beyond these conventional levels (e.g. 0.05, 0.01, 0.001). Later, however, Fisher (1956) recommended that fixed significance levels (e.g. 0.05) were too restrictive and argued that a researcher's significance level would depend on circumstances. Fisher also introduced the idea of fiducial inference, although this approach is rarely used in the biological sciences – Mayo (1996) and Oakes (1986) provide details.

Jerzy Neyman and Egon Pearson (Neyman & Pearson 1928, 1933) offered a related but slightly different approach, which has sometimes been called statistical hypothesis testing. Their approach differed from Fisher's in a number of important ways (Oakes 1986, Royall 1997).

1. They argued that we should set a level of significance (e.g. 0.05) in advance of the data collection and stick with it – this is sometimes called fixed level testing. The significance level is interpreted as the proportion of times the H_0 would be wrongly rejected using this decision rule if the experiment were repeated many times and the H_0 was actually true. Under the Neyman–Pearson scheme, the P value provides no additional information beyond indicating whether we should reject the H_0 at our specified significance level (Oakes 1986). They emphasized making a dichotomous decision about the H_0 (reject or nor reject) and the possible errors associated with that decision (see below) whereas Fisher was more concerned with measuring evidence against the H_0. Whether P values provide a suitable measure of evidence is a matter of debate (e.g. Royall 1997) that we will consider further in Section 3.6.

2. Another major difference between the Fisher and the Neyman–Pearson approaches was that Neyman and Pearson explicitly incorporated an alternative hypothesis (H_A) into their scheme. The H_A is the alternative hypothesis that must be true if the H_0 is false, e.g. if the H_0 is that two

population means are equal, then the H_A is that they are different by some amount. In contrast, Fisher strongly opposed the idea of H_A in significance testing (Cohen 1990).

3. Neyman and Pearson developed the concepts of Type I error (long-run probability of falsely rejecting H_0, which we denote α) and Type II error (long-run probability of falsely not rejecting H_0, which we denote β) and their *a priori* significance level (e.g. $\alpha = 0.05$) was the long-run probability of a Type I error (Gigerenzer 1993). This led naturally to the concept of power (the probability of correctly rejecting a false H_0). Fisher strongly disagreed with Neyman and Pearson about the relevance of the two types of error and even criticized Neyman and Pearson for having no familiarity with practical application of hypothesis testing in the natural sciences (Oakes 1986)!

Statisticians have recently revisited the controversy between the Fisher and Neyman–Pearson approaches to hypothesis testing (Inman 1994, Lehmann 1993, Mulaik *et al.* 1997, Royall 1997), pointing out their similarities as well as their disagreements and the confusion in terminology. Biologists, like psychologists (Gigerenzer 1993), most commonly follow a hybrid approach, combining aspects of both Fisherian inference and Neyman–Pearson decision-making to statistical hypothesis testing.

1. Specify H_0, H_A and appropriate test statistic

2. Specify *a priori* significance level (e.g. 0.05), which is the long-run frequency of Type I errors (α) we are willing to accept.

3. Collect the data by one or more random samples from the population(s) and calculate the test statistic from our sample data.

4. Compare that value of the statistic to its sampling distribution, assuming H_0 true.

5. If the probability of obtaining this value or one greater is less than the specified significance level (e.g. 0.05), then conclude that the H_0 is false and reject it ("significant" result),

6. If the probability of obtaining this value is greater than or equal to the specified significance level (e.g. 0.05), then conclude there is no evidence that the H_0 is false and retain it ("non-significant" result).

The Fisherian aspect of this hybrid approach is that some biologists use $P < 0.05$ (significant), $P < 0.01$ (very significant) and $P < 0.001$ (highly significant) or present the actual P values to indicate strength of evidence against the H_0. Although the latter has been strongly criticized by some in the psychological literature (Shaver 1993), there is some logical justification for providing P values (Oakes 1986). For one thing, it allows readers to use their own *a priori* significance levels to decide whether or not to reject the H_0.

To reiterate, interpretations from classical statistical tests are based on a long-run frequency interpretation of probabilities, i.e. the probability in a long run of identical "trials" or "experiments". This implies that we have one or more clearly defined population(s) from which we are sampling and for which inferences are to be made. If there is no definable population from which random samples are collected, the inferential statistics discussed here are more difficult to interpret since they are based on long-run frequencies of occurrence from repeated sampling. Randomization tests (Section 3.3.2), which do not require random sampling from a population, may be more applicable.

3.1.2 Associated probability and Type I error

Fisher and Neyman & Pearson both acknowledged that probabilities from classical statistical hypothesis testing must be interpreted in the long-run frequency sense, although the latter were more dogmatic about it. The sampling distribution of the test statistic (e.g. t) gives us the long-run probabilities of different ranges of t values occurring if we sample repeatedly from a population(s) in which the H_0 is true. The P value, termed the associated probability by Oakes (1986), then is simply the long-run probability of obtaining our sample test statistic *or one more extreme*, if H_0 is true. Therefore, the P value can be expressed as $P(\text{data}|H_0)$, the probability of observing our sample data, or data more extreme, under repeated identical experiments if the H_0 is true. This is not the same as the probability of H_0 being true, given the observed data – $P(H_0|\text{data})$. As Oakes (1986) has pointed out, there is rarely a sensible long-run frequency interpretation for the

probability that a particular hypothesis is true. If we wish to know the probability of H_0 being true, we need to tackle hypothesis testing from a Bayesian perspective (Berger & Berry 1988; see Section 3.7).

The *P* value is also sometimes misinterpreted as the probability of the result of a specific analysis being due to chance, e.g. a *P* value of <0.05 means that there is a less than 5% probability that the result is due to chance. This is not strictly correct (Shaver 1993); it is the probability of a result occurring by chance in the long run if H_0 is true, not the probability of any particular result being due to chance.

Traditionally, biologists are correctly taught that a non-significant result (not rejecting H_0) does not indicate that H_0 is true, as Fisher himself stressed. In contrast, the Neyman–Pearson logic is that H_0 and H_A are the only alternatives and the non-rejection of H_0 implies the acceptance of H_0 (Gigerenzer 1993), a position apparently adopted by some textbooks, e.g. Sokal & Rohlf (1995) refer to the acceptance of H_0. The Neyman–Pearson approach is really about alternative courses of actions based on the decision to accept or reject. Accepting the H_0 does not imply its truth, just that one would take the action that results from such a decision.

Our view is that a statistically non-significant result basically means we should suspend judgement and we have no evidence to reject the H_0. The exception would be if we show that the power of our test to detect a desired alternative hypothesis was high, then we can conclude the true effect is probably less than this specific effect size (Chapter 7). Underwood (1990, 1999) has argued that retention of the H_0 implies that the research hypothesis and model on which it is based are falsified (see Chapter 1). In this context, a statistically non-significant result should initiate a process of revising or even replacing the model and devising new tests of the new model(s). The philosophical basis for interpreting so-called 'negative' results continues to be debated in the scientific literature (e.g. see opinion articles by Allchin 1999, Hull 1999 and Ruse 1999 in *Marine Ecology Progress Series*).

The Type I error rate is the long-run probability of rejecting the H_0 at our chosen significance level, e.g. 0.05, if the H_0 is actually true in all the repeated experiments or trials. A Type I error is one of the two possible errors when we make a decision about whether the H_0 is likely to be true or not under the Neyman–Pearson protocol. We will consider these errors further in Section 3.2.

3.1.3 Hypothesis tests for a single population

We will illustrate testing an H_0 with the simplest type of test, the single-parameter *t* test. We demonstrated the importance of the *t* distribution for determining confidence intervals in Chapter 2. It can also be used for testing hypotheses about single population parameters or about the difference between two population parameters if certain assumptions about the variable hold. Here we will look at the first type of hypothesis, e.g. does the population mean equal zero? The value of the parameter specified in the H_0 doesn't have to be zero, particularly when the parameter is a mean, e.g. testing an H_0 that the mean size of an organism is zero makes little biological sense. Sometimes testing an H_0 that the mean equals zero is relevant, e.g. the mean change from before to after a treatment equals zero, and testing whether other parameters equal zero (e.g. regression coefficients, variance components, etc.) is very important. We will consider these parameters in later chapters.

The general form of the *t* statistic is:

$$t_s = \frac{St - \theta}{S_{St}} \tag{3.2}$$

where *St* is the value of the statistic from our sample, θ is the population value against which the sample statistic is to be tested (as specified in the H_0) and S_{St} is the estimated standard error of the sample statistic. We will go through an example of a statistical test using a one-sample *t* test.

1. Specify the H_0 (e.g. $\mu = 0$) and H_A (e.g. $\mu \neq 0$).

2. Take a random sample from a clearly defined population.

3. Calculate $t = (\bar{y} - 0)/s_{\bar{y}}$ from the sample, where $s_{\bar{y}}$ is the estimated standard error of the sample mean. Note that if H_0 is true, we would expect *t* to be close to zero, i.e. when we sample from a population with a mean of zero, most

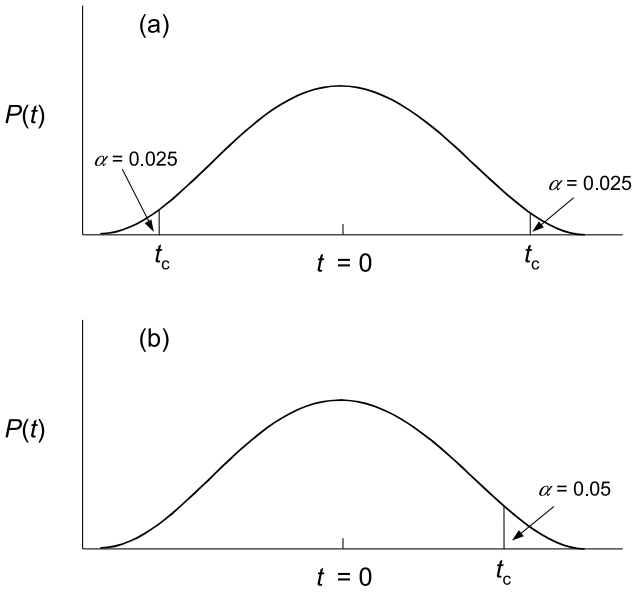

Figure 3.1 Probability distributions of t for (a) two-tailed and (b) one-tailed tests, showing critical t values (t_c).

samples will have means close to zero. Sample means further from zero are less likely to occur if H_0 is true. The probability of getting a sample mean a long way from zero, and therefore a large t, either positive or negative, is less if the H_0 is true. Large t values are possible if H_0 is true – they are just unlikely.

4. Compare t with the sampling distribution of t at $\alpha = 0.05$ (or 0.01 or whatever significance level you choose *a priori*) with $n-1$ df. Look at the t distribution in Figure 3.1. Values of t greater than $+t_c$ or less than $-t_c$ have a less than 0.05 chance of occurring from this t distribution, which is the probability distribution of t when H_0 is true. This value (t_c) is sometimes called the critical value. If the probability (P value) of obtaining our sample t value or one larger is less than 0.05 (our α), then we reject the H_0. Because we can reject H_0 in either direction, if μ is greater than zero or if μ is less than zero, then large values of the test statistic at either end of the sampling distribution will result in rejection of H_0 (Figure 3.1). This is termed a two-tailed test (see Section 3.1.4). To do a test with $\alpha = 0.05$, then we reject H_0 if our t value falls in the regions where $P = 0.025$ at each end of the sampling distribution ($0.025 + 0.025 = 0.05$). If the probability (P value) of obtaining our t value or one larger is ≥ 0.05, then we do not reject the H_0.

As mentioned earlier, the sampling distribution of the t statistic when the H_0 is true is also called the central t distribution. The probabilities for the t distribution for different degrees of freedom are tabled in most textbooks (usually for $P = 0.05$, 0.01 and sometimes 0.001). In addition, t distributions are programmed into statistical software. When using statistical tables, our value of t is simply compared to the critical t_c value at $\alpha = 0.05$. Larger t values always have a smaller P value (probability of this or a larger value occurring if H_0 is true) so if the statistic is larger than the critical value at 0.05, then H_0 is rejected. Statistical software usually gives actual P values for statistical tests, making the use of tables unnecessary.

We could theoretically use the sampling distribution of the sample mean (which would be a normal distribution) to test our H_0. However, there are an infinite number of possible combinations of mean and variance, so in practice such sampling distributions are not calculated. Instead, we convert the sample mean to a t value (subtracting μ specified in H_0 and dividing by the standard error of the mean), whose central distribution is well defined.

Finally, it is important to note the relationship between the hypothesis test illustrated here and confidence intervals described in Chapter 2. The H_0 that μ equals zero is tested using a t distribution; a confidence interval for μ is also constructed using the same t distribution (based on $n-1$ df). Not surprisingly then, a test of this H_0 with a 0.05 significance level is the equivalent of seeing whether the 95% (0.95) confidence interval for μ overlaps zero; if it does, we have no evidence to reject H_0.

3.1.4 One- and two-tailed tests

In most cases in biology, the H_0 is one of no effect (e.g. no difference between two means) and the H_A (the alternative hypothesis) can be in either direction; the H_0 is rejected if one mean is bigger than the other mean or vice versa. This is termed a two-tailed test because large values of the test statistic at either end of the sampling distribution will result in rejection of H_0 (Figure 3.1). The H_0 that a parameter equals a specific value is sometimes called a simple hypothesis or a point hypothesis (Barnett 1999). To do a test with $\alpha = 0.05$, then we use critical values of the test statistic at $\alpha = 0.025$ at each end of the sampling distribution. Sometimes, our H_0 is more specific than just no difference. We might only be interested in whether one mean is bigger than the other mean but not the other way. For example, we might expect increased density of organisms to induce competition and reduce their growth rate, and we can think of no mechanism whereby the organisms at the higher density would increase their growth. Here our H_0 is that the population mean growth rate for increased density is greater than or equal to the population mean growth rate for lower density. Our H_A is, therefore, that the population mean growth rate for increased density is less than the population mean growth rate for lower density. This is a one-tailed test, the H_0 being directional or composite (Barnett 1999), because only large values of the test statistic at one end of the sampling distribution will result in rejection of the H_0 (Figure 3.1). To do a test with $\alpha = 0.05$, then we use critical values of the test statistic at $\alpha = 0.05$ at one end of the sampling distribution.

We should test one-tailed hypotheses with care because we are obliged to ignore large differences in the other direction, no matter how tempting it may be to deal with them. For example, if we expect increased phosphorous (P) to increase plant growth compared to controls (C) with no added phosphorous, we might perform a one-tailed t test (H_0: $\mu_P \leq \mu_C$; H_A: $\mu_P > \mu_C$). However, we cannot draw any formal conclusions if growth rate is much less when phosphorous is added, only that it is a non-significant result and we have no evidence to reject the H_0. Is this unrealistic, expecting a biologist to ignore what might be an important effect just because it was in the opposite direction to that expected? This might seem like an argument against one-tailed tests, avoiding the problem by never ruling out interest in effects in both directions and always using two-tailed tests. Royall (1997) suggested that researchers who choose one-tailed tests should be trusted to use them correctly, although he used the problems associated with the one-tail versus two-tail choice as one of his arguments against statistical hypothesis testing and *P* values more generally. An example of one-tailed tests comes from Todd & Keough (1994), who were interested in whether microbial films that develop on marine hard substrata act as cues inducing invertebrate larvae to settle. Because they expected these films to be a positive cue, they were willing to focus on changes in settlement in one direction only. They then ignored differences in the opposite direction from their *a priori* one-tailed hypothesis.

Most statistical tables either provide critical values for both one- and two-tailed tests but some just have either one- or two-tailed critical values depending on the statistic, so make sure you look up the correct *P* value if you must use tables. Statistical software usually produces two-tailed *P* values so you should compare the *P* value to $\alpha = 0.10$ for a one-tailed test at 0.05.

3.1.5 Hypotheses for two populations

These are tests of null hypotheses about the equivalent parameter in two populations. These tests can be one- or two-tailed although testing a point null hypothesis with a two-tailed test is more common in practice, i.e. the parameter is the same in the two populations. If we have a random sample from each of two independent populations, i.e. the populations represent different collections of observations (i.e. sampling or experimental units), then to test the H_0 that $\mu_1 = \mu_2$ (comparing two independent population means):

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s_{\bar{y}_1 - \bar{y}_2}} \tag{3.3}$$

where

$$s_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{3.4}$$

Equation 3.4 is the standard error of the difference between the two means. This is just like the

one-parameter t test except the single sample statistic is replaced by the difference between two sample statistics, the population parameter specified in the H_0 is replaced by the difference between the parameters of the two populations specified in the H_0 and the standard error of the statistic is replaced by the standard error of the difference between two statistics:

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{s_{\bar{y}_1 - \bar{y}_2}} \tag{3.5}$$

We follow the steps in Section 3.1.1 and compare t to the t distribution with $n_1 + n_2 - 2$ df in the usual manner. This H_0 can also be tested with an ANOVA F-ratio test (Chapter 8).

We will illustrate tests of hypotheses about two populations with two examples. Ward & Quinn (1988) studied aspects of the ecology of the intertidal predatory gastropod *Lepsiella vinosa* on a rocky shore in southeastern Australia (Box 3.1). *L. vinosa* occurred in two distinct zones on this shore: a high-shore zone dominated by small grazing gastropods *Littorina* spp. and a mid-shore zone dominated by beds of the mussels *Xenostrobus pulex* and *Brachidontes rostratus*. Both gastropods and mussels are eaten by *L. vinosa*. Other data indicated that rates of energy consumption by *L. vinosa* were much greater in the mussel zone. Ward & Quinn (1988) were interested in whether there were any differences in fecundity of *L. vinosa*, especially the number of eggs per capsule, between the zones. From June to September 1982, they collected any egg capsules they could find in each zone and recorded the number of eggs per capsule. There were 37 capsules recorded from the littorinid zone and 42 from the mussel zone. The H_0 was that there is no difference between the zones in the mean number of eggs per capsule. This is an independent comparison because the egg capsules were independent between the zones.

Furness & Bryant (1996) studied energy budgets of breeding northern fulmars (*Fulmarus glacialis*) in Shetland (Box 3.2). As part of their study, they recorded various characteristics of individually labeled male and female fulmars. We will focus on differences between sexes in metabolic rate. There were eight males and six females labeled. The H_0 was that there is no difference between the sexes in the mean metabolic rate of fulmars. This is an independent comparison because individual fulmars can only be either male or female.

If we have a random sample from a population and we have recorded two (paired) variables from each observation, then we have what are commonly called paired samples, e.g. observations at two times. To test whether the population mean difference between the two sets of observations equals zero, we basically use a test for a single population (Section 3.1.3) to test the H_0 that $\mu_d = 0$:

$$t = \frac{\bar{d}}{s_{\bar{d}}} \tag{3.6}$$

where $\bar{d}$ is the mean of the pairwise differences and $s_{\bar{d}}$ is the standard error of the pairwise differences. We compare t with a t distribution with $n - 1$ df in the usual manner. This H_0 can also be tested with a two factor unreplicated ANOVA F-ratio test (Chapter 10).

For example, Elgar *et al*. (1996) studied the effect of lighting on the web structure of an orb-spinning spider (Box 3.3). They set up wooden frames with two different light regimes (controlled by black or white mosquito netting), light and dim. A total of 17 orb spiders were allowed to spin their webs in both a light frame and a dim frame, with six days' "rest" between trials for each spider, and the vertical and horizontal diameter of each web was measured. Whether each spider was allocated to a light or dim frame first was randomized. The null hypotheses were that the two variables (vertical diameter and horizontal diameter of the orb web) were the same in dim and light conditions. Elgar *et al*. (1996) correctly treated this as a paired comparison because the same spider spun her web in a light frame and a dark frame.

We can also test whether the variances of two populations are the same. Recall from Chapter 2 that variances are distributed as chi-squares and the ratio of two chi-square distributions is an F distribution, another probability distribution that is well defined. To test the H_0 that $\sigma_1^2 = \sigma_2^2$ (comparing two population variances), we calculate an F-ratio statistic:

$$F = \frac{s_1^2}{s_2^2} \tag{3.7}$$

Box 3.1 Fecundity of predatory gastropods

Ward & Quinn (1988) collected 37 egg capsules of the intertidal predatory gastropod *Lepsiella vinosa* from the littorinid zone on a rocky intertidal shore and 42 capsules from the mussel zone. Other data indicated that rates of energy consumption by *L. vinosa* were much greater in the mussel zone so there was interest in differences in fecundity between the zones. The H_0 was that there is no difference between the zones in the mean number of eggs per capsule. This is an independent comparison because individual egg capsules can only be in either of the two zones.

Zone	*n*	Mean	Median	Rank sum	Standard deviation	SE of mean	95% CI for mean
Littorinid	37	8.70	9	1007	2.03	0.33	8.03–9.38
Mussel	42	11.36	11	2153	2.33	0.36	10.64–12.08

Note that standard deviations (and therefore the variances) are similar and boxplots (Figure 4.4) do not suggest any asymmetry so a parametric *t* test is appropriate.

Pooled variance test:

$t = -5.39$, df = 77, $P < 0.001$.

We would reject the H_0 and conclude there was a statistically significant difference in mean number of eggs per capsule between zones.

Effect size (difference between means) = −2.65 (95% CI: −1.674 to −3.635)

Separate variance test:

$t = -5.44$, df = 77, $P < 0.001$.

Note that the *t* values were almost identical and the degrees of freedom were the same, not surprising since the variances were almost identical.

Although there was little justification for a non-parametric test, we also tested the H_0 that there was no difference in a more general measure of location using the Mann–Whitney–Wilcoxon test.

$U = 304.00$, χ^2 approximation = 21.99 with 1 df, $P < 0.001$.

Again we would reject the H_0. In this example, the parametric pooled and separate variance *t* tests and non-parametric test all give *P* values < 0.001.

A randomization test was done to test the H_0 that there is no difference between the mean number of eggs per capsule so that any possible allocation of observations to the two groups is equally likely.

Mean difference = −2.65, $P < 0.001$ (significant) for difference as or more extreme than observed based on 10 000 randomizations.

where s_1^2 is the larger sample variance and s_2^2 is the smaller sample variance. We compare this *F*-ratio with an *F* distribution with $n_1 - 1$ df for numerator (sample one) and $n_2 - 1$ df for denominator (sample two). We will consider *F*-ratio tests on variances in more detail in Chapters 5 onwards.

3.1.6 Parametric tests and their assumptions

The *t* tests we have just described for testing null hypotheses about population means are classified as parametric tests, where we can specify a probability distribution for the populations of the

Box 3.2 | Metabolic rate of male and female fulmars

Furness & Bryant (1996) studied energy budgets of breeding northern fulmars (*Fulmarus glacialis*) in Shetland. As part of their study, they recorded various characteristics of individually labeled male and female fulmars. We will focus on differences between sexes in metabolic rate. There were eight males and six females labeled. The H_0 was that there is no difference between the sexes in the mean metabolic rates of fulmars. This is an independent comparison because individual fulmars can only be either male or female.

Sex	*n*	Mean	Median	Standard deviation	SE of mean	95% CI for mean
Male	8	1563.78	1570.55	894.37	316.21	816.06 – 2311.49
Female	6	1285.52	1226.15	420.96	171.86	843.74 – 1727.29

Note that variances are very different although the boxplots (Figure 4.5) do not suggest strong asymmetry. The small and unequal sample sizes, in conjunction with the unequal variances, indicate that a *t* test based on separate variances is more appropriate.

Separate variance test:

$t = 0.77$, df = 10.5, $P = 0.457$.

We would not reject the H_0 and conclude there was no statistically significant difference in mean metabolic rate of fulmars between sexes.

The effect size (difference between means) = 278.26 (95% CI: −518.804 to 1075.321).

Note that the confidence interval on the mean difference includes zero, as expected given the non-significant result from the test.

The very different variances would make us reluctant to use a rank-based non-parametric test. Even a randomization test might be susceptible to unequal variance, although the results from such a test support the previous conclusion.

Mean difference = 278.26, $P = 0.252$ (not significant) for difference as or more extreme than observed based on 10 000 randomizations.

variable from which our samples came. All statistical tests have some assumptions (yes, even so-called "non-parametric tests" – see Section 3.3.3) and if these assumptions are not met, then the test may not be reliable. Basically, violation of these assumptions means that the test statistic (e.g. *t*) may no longer be distributed as a *t* distribution, which then means that our *P* values may not be reliable. Although parametric tests have these assumptions in theory, in practice these tests may be robust to moderate violations of these assumptions, i.e. the test and the *P* values may still be reliable even if the assumptions are not met. We will describe the assumptions of *t* tests here and introduce ways of checking these assumptions, although these methods are presented in more detail in Chapter 4. The assumptions themselves are also considered in more detail as assumptions for linear models in Chapters 5 onwards.

The first assumption is that the samples are from normally distributed populations. There is reasonable evidence from simulation studies (Glass *et al.* 1972, Posten 1984) that significance tests based on the *t* test are usually robust to violations of this assumption unless the distributions are very non-symmetrical, e.g. skewed or multimodal. Checks for symmetry of distributions can include dotplots (if *n* is large enough), boxplots and

Box 3.3 | Orb spider webs and light intensity

Elgar *et al.* (1996) exposed 17 orb spiders each to dim and light conditions and recorded two aspects of web structure under each condition. The H_0s are that the two variables (vertical diameter and horizontal diameter of the orb web) were the same in dim and light conditions. Because the same spider spun her web in both light conditions, then this was a paired comparison. Boxplots of paired differences for both variables suggested symmetrical distributions with no outliers, so a parametric paired *t* test is appropriate.

Horizontal diameter (cm):

Mean difference = 46.18, SE difference = 21.49.
$t = 2.15$, df = 16, $P = 0.047$ (significant).

So we would reject the H_0 and conclude that, for the population of female orb spiders, there is a difference in the mean horizontal diameter of spider webs between light and dim conditions.

Wilcoxon signed rank $z = -1.84$, $P = 0.066$ (not significant), do not reject H_0. Note the less powerful non-parametric test produced a different result.

Vertical diameter (cm):

Mean difference = 20.59, SE difference = 21.32.
$t = 0.97$, df = 16, $P = 0.349$ (not significant), do not reject H_0.

So we would not reject the H_0 and conclude that, for the population of female orb spiders, there is no difference in the mean vertical diameter of spider webs between light and dim conditions.

Wilcoxon signed rank $z = -0.78$, $P = 0.434$ (not significant). In this case, the non-parametric test produced the same conclusion as the *t* test.

pplots (see Chapter 4). Transformations of the variable to a different scale of measurement (Chapter 4) can often improve its normality. We do not recommend formal significance tests for normality (e.g. Shapiro–Wilk test, Lilliefors test; see Sprent 1993) because, depending on the sample size, these tests may reject the H_0 of normality in situations when the subsequent *t* test may be reliable.

The second assumption is that samples are from populations with equal variances. This is a more critical assumption although, again, the usual *t* test is very robust to moderately unequal variances if sample sizes are equal (Glass *et al.* 1972, Posten 1984). While much of the simulation work relates to analysis of variance (ANOVA) problems (see Day & Quinn 1989, Wilcox *et al.* 1986, Chapter 8), the results also hold for *t* tests, which are equivalent to an ANOVA *F*-ratio test on two groups. For example, if *n* equals six and the ratio of the two standard deviations is four or less, simulations show that the observed Type I error rate for the *t* test is close to the specified rate (Coombs *et al.* 1996). If sample sizes are very unequal, especially if the smaller sample has the larger variance, then Type I error rates may be much higher than postulated significance level. If the larger sample has the larger variance, then the rate of Type II errors will be high (Judd *et al.* 1995, Coombs *et al.* 1996). Coombs *et al.* (1996) illustrated this with simulation data from Wilcox *et al.* (1986) that showed that for sample sizes of 11 and 21, a four to one ratio of standard deviations (largest standard deviation associated with small sample size) resulted in a Type I error rate of nearly 0.16 for a nominal α of 0.05. Note that unequal variances are often due to skewed distributions, so fixing the non-normality problem will often make variances more similar. Checks for this assumption include

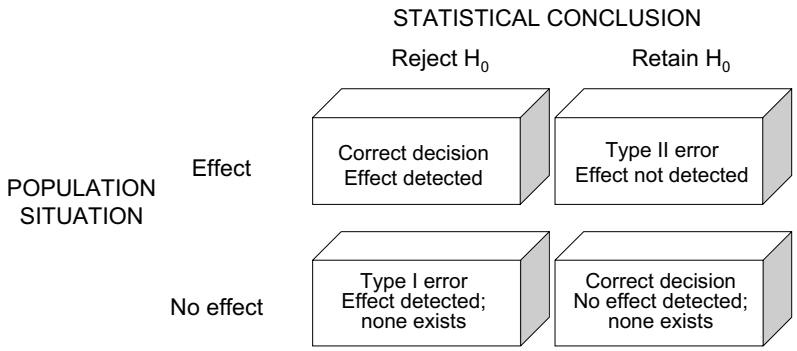

Figure 3.2 Statistical decisions and errors when testing null hypotheses.

examining boxplots of each sample for similar spreads. We do not routinely recommend a preliminary test of equal population variances using an *F*-ratio test (Section 3.1.5) for three reasons.

- The *F*-ratio test might be more sensitive to non-normality than the *t* test it is "protecting".
- Depending on sample size, an *F*-ratio test may not detect variance differences that could invalidate the following *t* test, or it might find unequal variances (and hence recommend the following analysis not be done), which would not adversely affect the subsequent *t* test (Markowski & Markowski 1990). This dependence of the results of a statistical hypothesis test on sample size is well known and will be discussed further in Section 3.6.
- Statistical hypothesis testing should be used carefully, preferably in situations where power and effect sizes have been considered; this is rarely the case for exploratory checks of assumptions.

The third assumption is that the observations are sampled randomly from clearly defined populations. This is an assumption that must be considered at the design stage. If samples cannot be sampled randomly from populations, then a more general hypothesis about differences between samples can be tested with a randomization test (see Section 3.3.2).

These *t* tests are much more sensitive to assumptions about normality and equal variances if sample sizes are unequal, so for this reason alone, it's always a good idea to design studies with equal sample sizes. On an historical note, testing differences between means when the variances also differ has been a research area of long-standing interest in statistics and is usually called the Behrens–Fisher problem. Solutions to this problem will be discussed in Section 3.3.1.

An additional issue with many statistical tests, including parametric tests, is the presence of outliers (Chapter 4). Outliers are extreme values in a sample very different from the rest of the observations and can have strong effects on the results of most statistical tests, in terms of both Type I and Type II errors. Note that both parametric *t* tests and non-parametric tests based on ranks (Section 3.3) are affected by outliers (Zimmerman 1994), although rank-based tests are less sensitive (Zimmerman & Zumbo 1993). Detection and treatment of outliers is considered in Chapter 4.

3.2 Decision errors

3.2.1 Type I and II errors

When we use the Neyman–Pearson protocol to test an H_0, there are four possible outcomes based on whether the H_0 was actually true (no effect) or not (real effect) for the population (Figure 3.2). A rejection of a H_0 is usually termed a significant result (statistically significant, not necessarily biologically significant – see Box 3.4) and implies that some alternative hypothesis (H_A) is true. Clearly, two of the outcomes result in the right statistical decision being made; we correctly reject a false H_0 or we correctly retain a true H_0. What about the two errors?

- A Type I error is when we mistakenly reject a correct H_0 (e.g. when we conclude from our sample and a *t* test that the population parameter is not equal to zero when in fact the population parameter does equal zero) and is denoted α. A Type I error can only occur when H_0 is true.
- A Type II error is when we mistakenly accept an incorrect H_0 (e.g. when we conclude from

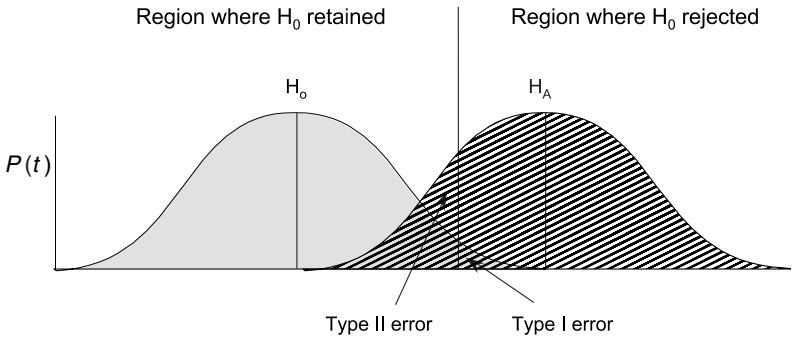

Figure 3.3 Graphical representation of Type I and Type II error probabilities, using a *t* test as an example.

our sample and a *t* test that the population parameter equals zero when in fact the population parameter is different from zero). Type II error rates are denoted by β and can only occur when the H_0 is false.

Both errors are the result of chance. Our random sample(s) may provide misleading information about the population(s), especially if the sample sizes are small. For example, two populations may have the same mean value but our sample from one population may, by chance, contain all large values and our sample from the other population may, by chance, contain all small values, resulting in a statistically significant difference between means. Such a Type I error is possible even if H_0 ($\mu_1 = \mu_2$) is true, it's just unlikely. Keep in mind the frequency interpretation of *P* values also applies to the interpretation of error rates. The Type I and Type II error probabilities do not necessarily apply to our specific statistical test but represent the long-run probability of errors if we repeatedly sampled from the same population(s) and did the test many times.

Examine Figure 3.3, which shows the probability sampling distribution of *t* when the H_0 is true (left curve) and the probability sampling distribution of *t* when a particular H_A is true (right curve). Of course, we never know what this latter distribution looks like in practice because if H_0 is false, we don't know what the real H_A is. For a particular df, there will be a different distribution for each possible H_A but only one sampling distribution for H_0. The critical value of *t* for $\alpha = 0.05$ is indicated. If H_0 is actually true, any *t* value greater than this critical value will lead to a rejection of H_0 and a Type I error. If H_0 is actually false and H_A is true, any value equal to or smaller than this critical value will lead to non-rejection of H_0 and a Type II error. Note that if H_0 is, for example, no difference between means, then H_A is a difference between means. The bigger the difference, the further the *t* distribution for H_A will be to the right of the *t* distribution for H_0 and the less likely will be a Type II error.

Traditionally, scientists have been most concerned with Type I errors. This is probably because statistically significant results imply falsification of a null hypothesis and therefore progress in science and maybe because we wrongly equate statistical significance with biological significance (see Box 3.4). Therefore, we protect ourselves (and our discipline) from false significant results by using a conservative significance level (e.g. 0.05); this means that we are controlling our Type I error rate to 0.05 or 5%. If the probability of obtaining our sample when the H_0 is true is less than 0.05, then we reject that H_0; otherwise we don't reject it. Why don't we use an even lower significance level to protect ourselves from Type I errors even more? Mainly because for most statistical tests, for a given sample size and level of variation, lowering the Type I error rate (the significance level) results in more Type II errors (imagine moving the vertical line to the right in Figure 3.3) if it turns out that the H_A is true.

For some activities, especially environmental monitoring and impact assessment and experiments involving human health issues, Type II errors may be of much greater importance than Type I. Consider a monitoring program, and the consequences of the two kinds of errors. A Type I error results in an erroneous claim of a significant environmental change. In an ideal world, the result would be a requirement by the relevant regulatory authority for some mitigation or cessation of the activity causing that change. The

Box 3.4 Biological versus statistical significance

It is important to distinguish between biological and statistical significance. As mentioned in Section 3.6.1, if we take larger and larger samples, we can detect even very small differences. Whenever we get a (statistically) significant result, we must still decide whether the effects that we observe are biologically meaningful. For example, we might measure 100 snails in each of two populations, and we would almost certainly find that the two populations were different in size. However, if the mean size differed by ≈1%, we may struggle to explain the biological meaning of such a small difference.

What is biologically significant? The answer has nothing to do with statistics, but with our biological judgment, and the answer will vary with the questions being answered. Small effects of experimental treatments may be biologically significant when we are dealing with rates of gene flow, selection, or some physiological measurements, because small differences can have important repercussions in population genetics or organism health. For example, small changes in the concentration of a toxin in body tissues may be enough to cause mortality. In contrast, small effects may be less important for ecological processes at larger spatial scales, especially under field conditions.

It is important for biologists to think carefully about how large an effect has to be before it is biologically meaningful. In particular, setting biologically important effect sizes is crucial for ensuring that out statistical test has adequate power.

"costs" would be purely financial – the cost of (unnecessary) mitigation. A Type II error, on the other hand, is a failure to detect a change that has occurred. The verdict of "no significant impact" results in continuation of harmful activities. There is no added financial cost, but some time in the future the environmental change will become large enough to become apparent. The consequence of this error is that significant environmental degradation may have occurred or become more widespread than if it had been detected early, and mitigation or rehabilitation may be necessary, perhaps at significant cost. A strong argument can therefore be made that for many "applied" purposes, Type II errors are more important than Type I errors. A similar argument applies to other research areas. Underwood (1990, 1997), in describing the logical structure of hypothesis testing, indicates very clearly how Type II errors can misdirect research programs completely.

The inverse of Type II error is power, the probability of rejecting a false H_0. We will consider power in more detail as part of experimental design in Chapter 7.

3.2.2 Asymmetry and scalable decision criteria

One of the problems of fixing our significance level α, even if we then use power analysis to determine sample sizes to minimize the probability of Type II errors, is that there is an implicit asymmetry in the importance of H_0 relative to H_A (Barnett 1999, Oakes 1986). In many practical situations, fixing α to 0.05 will make it difficult to reduce the probability of Type II errors to a comparable level, unless sample sizes or effect sizes are very large. The only solution to this problem, while still maintaining the structure of statistical tests and errors associated with decisions, is to abandon fixed level testing and use decision criteria that provide a more sensible balance between Type I and Type II errors.

Mapstone (1995) has proposed one way of incorporating flexible decision criteria in statistical hypothesis testing in ecology and environmental science. He suggested that we should set the ratio of acceptable Type I and Type II errors *a priori*, based on the relative costs of making each kind of error, and the critical effect size is the most crucial element. Keough & Mapstone (1995)

have incorporated this idea into a framework for designing environmental monitoring programs, and included a worked example. Downes *et al.* (2001) have also advocated scalable decision criteria for assessing environmental impact in freshwater ecosystems. The logic of considering costs of making errors in statistical decision making is much closer to the Bayesian approach to making decisions, although Bayesians eschew the long-run frequency view of probability (Section 3.7).

3.3 Other testing methods

The statistical tests most commonly used by biologists, and the tests based on the *t* distribution we have just described, are known as parametric tests. These tests make distributional assumptions about the data, which for *t* tests are that the distributions of the populations from which the samples came are normal. Most textbooks state that parametric tests are robust to this assumption, i.e. the sampling distribution of the *t* statistic still follows the appropriate mathematical distribution even if the variable has a non-normal distribution. This means that the conclusions from the test of H_0 are still reliable even if the underlying distribution is not perfectly normal. This robustness is limited, however, and the assumption of normality (along with other assumptions inherent in all statistical tests – see Section 3.1.6) should always be checked before doing a parametric analysis.

3.3.1 Robust parametric tests

A number of tests have been developed for the H_0 that $\mu_1 = \mu_2$ which do not assume equal variances. For example, there are approximate versions of the *t* test (called variously the Welch test, Welch–Aspin test, the Satterthwaite-adjusted *t* test, Behrens–Fisher test, separate variances *t* test), which are available in most statistical software. The most common version of this test recalculates the df for the *t* test as (Hays 1994):

$$\frac{(s_1/\sqrt{n_1} + s_2/\sqrt{n_2})^2}{(s_1/\sqrt{n_1})^2/(n_1+1) + (s_2/\sqrt{n_2})^2/(n_2+1)} - 2 \quad (3.8)$$

This results in lower df (which may not be an integer) and therefore a more conservative test. Such a test is more reliable than the traditional *t* test when variances are very unequal and/or sample sizes are unequal.

Coombs *et al.* (1996) reviewed all the available tests for comparing two population means when variances may be unequal. They indicated that the Welch test is suitable when the samples come from normally distributed populations but recommended the Wilcox *H* test, based on M-estimators and bootstrapped estimates of variance (Chapter 2), for skewed distributions. Unfortunately, this test is not available in most software.

Some common types of null hypotheses can also be tested with non-parametric tests. Non-parametric tests do not assume that the underlying distribution of the population(s) from which the samples came is normal. Before looking at "classical" non-parametric tests based on ranks, let's consider another type of statistical test called a randomization test.

3.3.2 Randomization (permutation) tests

These tests resample or reshuffle the original data many times to generate the sampling distribution of a test statistic directly. Fisher (1935) first proposed that this method might be suitable for testing hypotheses but, without computers, could only analyze very small data sets. To illustrate randomization tests, we will revisit the example described in Section 3.1.5 where Ward & Quinn (1988) wished to test the H_0 that there is no difference between the mussel and littorinid zones in the mean number of eggs per capsule of *L.vinosa*. The steps in the randomization test are as follows (Manly 1997).

1. Calculate the difference between the mean numbers of eggs per capsule of the two groups (D_0).

2. Randomly reassign the 79 observations so that 37 are in the littorinid zone group and 42 are in the mussel zone group and calculate the difference between the means of the two groups (D_1).

3. Repeat this step a large number of times, each time calculating the D_i. How many randomizations? Manly (1997) suggested 1000 times for a 0.05 test and 5000 times for a

0.01 test. With modern computer power, these numbers of randomizations only take a few seconds.

4. Calculate the proportion of all the D_is that are greater than or equal to D_0 (the difference between the means in our samples). This is the "*P* value" and it can be compared to an *a priori* significance level (e.g. 0.05) to decide whether to reject the H_0 or not (Neyman–Pearson tradition), or used as a measure of "strength of evidence" against the H_0 (Fisher tradition – see Manly 1997).

The underlying principle behind randomization tests is that if the null hypothesis is true, then any random arrangement of observations to groups is equally possible (Crowley 1992). Randomization tests can be applied to situations where we are comparing groups or testing whether a set of observations occurs in a random order (e.g. time series). They are particularly useful when analyzing data for which the distribution is unknown (Potvin & Roff 1993), when random sampling from populations is not possible (e.g. we are using data that occurred opportunistically, such as museum specimens – see Manly 1997) or perhaps when other assumptions such as independence of observations are questionable, as when testing for temporal trends (Manly 1997). There are some potential interpretation problems with randomization tests that users should be aware of. First, they involve resampling the data to generate a probability distribution of the test statistic. This means that their results are more difficult to relate to any larger population but the positive side is that they are particularly useful for analyzing experiments where random sampling is not possible but randomization of observations to groups is used (Ludbrook & Dudley 1998). Crowley (1992, p. 432) argued that the difficulty of making inferences to some population is a problem "of greater theoretical than applied relevance" (see also Edgington 1995), particularly as randomization tests give similar *P* values to standard parametric tests when assumptions hold (Manly 1997). Manly (1997) also did not see this as a serious problem and pointed out that one of the big advantages of randomization tests is in situations when a population is not relevant or the whole population is effectively measured. Second, the H_0 being tested then is not one about population parameters, but simply that there is no difference between the means of the two groups, i.e. is the difference between group means "greater then we would expect by chance". Finally, the *P* value is interpreted differently from the usual "classical" tests. In randomization tests, the *P* value is the proportion of possible data rearrangements (e.g. between two groups) that are equal to, or more extreme than, the one we observed in our sample(s). Interestingly, because the *P* value is determined by a (re)sampling process, confidence intervals for the *P* value can be determined (Crowley 1992).

Randomization tests for differences between group means are not free of assumptions. For example, randomization tests of the H_0 of no difference between means are likely to be sensitive to differences in variances (Boik 1987, Stewart-Oaten *et al.* 1992). Indeed, randomization tests of location (e.g. mean) differences should be considered to have an assumption of similar distributions in the different samples, and transformations used where appropriate (Crowley 1992). So these tests should not be automatically applied to overcome problems of variance heterogeneity.

Manly (1997) is an excellent introduction to randomization tests from a biological perspective and Crowley (1992) critically summarized many applications of randomization tests in biology. Other good references for randomization tests are Edgington (1995) and Noreen (1989).

3.3.3 Rank-based non-parametric tests

Statisticians have appreciated the logic behind randomization tests for quite a long time, but the computations involved were prohibitive without computers. One early solution to this problem was to rank the observations first and then randomize the ranks to develop probability distributions of a rank-based test statistic. Ranking the observations has two advantages in this situation. First, determining the probability distribution of a rank-based test statistic (e.g. sum of the ranks in each sample) by randomization is relatively easy, because for a given sample size with no ties, the distribution is identical for any set of data. The critical values for such distributions are tabled in

many statistics books. In contrast, determining the probability distribution for a test statistic (e.g. difference between means) based on randomizing the original observations was not possible before computers except for small sample sizes. Second, using the ranks of the observations removes the assumption of normality of the underlying distribution(s) in each group, although other assumptions may still apply.

Although there is a wide range of rank-based non-parametric tests (Hollander & Wolfe 1999, Siegel & Castellan 1988, Sprent 1993), we will only consider two here. First, consider a test about differences between two populations. The Mann–Whitney–Wilcoxon test is actually two independently developed tests (Mann–Whitney and Wilcoxon) that produce identical results. The H_0 being tested is that the two samples come from populations with identical distributions against the H_A that the samples come from populations which differ only in location (mean or median). The procedure is as follows.

1. Rank all the observations, ignoring the groups. Tied observations get the average of their ranks.

2. Calculate the sum of the ranks for both samples. If the H_0 is true, we would expect a similar mixture of ranks in both samples (Sprent 1993).

3. Compare the smaller rank sum to the probability distribution of rank sums, based on repeated randomization of observations to groups, and test in the usual manner.

4. For larger sample sizes, the probability distribution of rank sums approximates a normal distribution and the *z* statistic can be used. Note that different software can produce quite different results depending on whether the large-sample approximation or exact randomization methods are used, and also how ties are handled (Bergmann *et al.* 2000).

Second, we may have a test about differences based on paired observations. For paired samples, we can use the Wilcoxon signed-rank test to test the H_0 that the two sets of observations come from the same population against the H_A that the populations differ in location (mean or median). This test is actually a test of a single population parameter, analyzing the paired differences, and the procedure is as follows.

1. Calculate the difference between the observations for each pair, noting the sign of each difference. If H_0 is true, we would expect roughly equal numbers of + and − signs.

2. Calculate the sum of the positive ranks and the sum of the negative ranks.

3. Compare the smaller of these rank sums to the probability distribution of rank sums, based on randomization, and test in the usual manner.

4. For larger sample sizes, the probability distribution of rank sums follows a normal distribution and the *z* statistic can be used, although the concern of Bergmann *et al.* (2000) about differences between the large sample approximation and exact methods for the Mann–Whitney–Wilcoxon test may also apply to the Wilcoxon signed-rank test.

Another non-parametric approach using ranks is the class of rank transformation tests. This is a more general approach that theoretically can be applied to any analysis for which there is a parametric test. The data are transformed to ranks and then these ranks are analyzed using the appropriate parametric analysis. Note that this technique is conceptually no different to transforming data to logs to meet the assumptions of a parametric test (Chapter 4) and is therefore not a true non-parametric test (Potvin & Roff 1993). The rank transform approach will generally give the same answer as the appropriate rank-based test, e.g. rank transform *t* test is the same as the Mann–Whitney–Wilcoxon test (Zimmerman & Zumbo 1993), although if there are a large number of ties the results will vary a little. Tests based on the rank transform method have also been used for various linear model analyses (Chapters 5, 8 and 9).

Although these non-parametric tests of location differences do not assume a particular shape (e.g. normal) of the underlying distributions, they do assume that the distributions of the populations are similar, so the assumption of equal variances still applies (Crowley 1992, Manly 1997, Sprent 1993, Stewart-Oaten *et al.* 1992, Zimmerman & Zumbo 1993). The common strategy in biological research to use rank-based

non-parametric tests to overcome variance heterogeneity is inappropriate. Variance heterogeneity in the two-sample hypothesis test should be dealt with by using a robust test, such as the Welch *t* test (Section 3.3.1) or by transforming the data to remove the relationship between the mean and variance (Chapter 4).

These non-parametric tests generally have lower power than the analogous parametric tests when parametric assumptions are met, although the difference in power is surprisingly small (e.g. <5% difference for Mann–Whitney–Wilcoxon test versus *t* test) given the former's use of ranks rather than the original data (Hollander & Wolfe 1999). With non-normal distributions, the non-parametric tests do cope better but because normality by itself is the least critical of all parametric assumptions, its hard to recommend the rank-based tests except in situations where (i) the distributions are very weird, and transformations do not help, or (ii) outliers are present (see Chapter 4). It is sometimes recommended that if the data are not measured on a continuous scale (i.e. the data are already in the form of ranks), then tests like the Mann–Whitney–Wilcoxon are applicable. We disagree because such a test is equivalent to applying a parametric test (e.g. *t* test) to the ranks, a much simpler and more consistent approach. It is also worth noting that the rank-based randomization tests don't really have any advantage over randomization tests based on the original data, except in terms of computation (which is irrelevant with modern computer power) – see Ludbrook & Dudley (1998). Both have assumptions of equal distributions in the two groups, and therefore equal variances, and neither is very sensitive to non-normality.

Rank-based tests have been argued to be more powerful than parametric tests for very skewed (heavy tailed) distributions. However, this is primarily because rank-based tests deal with outliers more effectively (Zimmerman & Zumbo 1993). Indeed, outliers cause major problems for parametric tests and their identification should be a priority for exploratory data analysis (Chapter 4). The alternative to rank-based tests is to remove or modify the outlying values by trimming or winsorizing (Chapter 2) and using a parametric test. Note that non-parametric tests are not immune to outliers; they are just not affected as much as parametric tests (Zimmerman & Zumbo 1993).

3.4 Multiple testing

3.4.1 The problem

One of the most difficult issues related to statistical hypothesis testing is the potential accumulation of decision errors under circumstances of multiple testing. As the number of tests increases, so does the probability of making at least one Type I error among the collection of tests. The probability of making one or more Type I errors in a set (or family) of tests is called the family-wise Type I error rate, although Day & Quinn (1989) and others have termed it experiment-wise Type I error rate because it is often used in the context of multiple comparisons of means when analyzing experimental data. The problem of increasing family-wise Type I error rate potentially occurs in any situation where there are multiple significance tests that are considered simultaneously. These include pairwise comparisons of treatment groups in an experiment (Chapter 8), testing pairwise correlations between multiple variables recorded from the same experimental or sampling units (Rice 1989) or multiple univariate analyses (e.g. *t* tests) of these variables.

If the tests are orthogonal (i.e. independent of each other), the family-wise Type I error can be calculated:

$$1-(1-\alpha)^c \tag{3.9}$$

where α is the significance level (e.g. 0.05) for each test and c is the number of tests. For example, imagine having a random sample from a number of populations and we wish to test the H_0s that each independent pair of population means is equal. We keep these comparisons independent by not using the same population in more than one test. As the number of populations we wish to compare increases, so does the number of pairwise comparisons required and the probability of at least one Type I error among the family of tests (Table 3.1). If the tests are non-orthogonal, then the family-wise Type I error rate will be lower (Ramsey 1993), but cannot be calculated as it will

Table 3.1 | Accumulation of probability of at least one Type 1 error among a "family" of tests

No. of tests	Family-wise probability of at least one Type I error
3	0.14
10	0.40
45	0.90

depend on the degree of non-independence among the tests.

The different approaches for dealing with the increased probability of a Type I error in multiple testing situations are based on how the Type I error rate for each test (the comparison-wise Type I error rate) is reduced to keep the family-wise Type I error rate at some reasonable level. Each test will then have a more stringent significance level but as a consequence, much reduced power if the H_0 is false. However, the traditional priority of recommendations for dealing with multiple testing has been strict control of family-wise Type I error rates rather than power considerations. Before describing the approaches for reducing the Type I error rate for each test to control the family-wise Type I error rate, we need to consider two other issues. The first is how we define the family of tests across which we wish to control the Type I error rate and the second is to what level should we control this error rate.

What comprises a family of tests (Shaffer 1995, Hancock & Klockars 1996) for determining error rates is a difficult decision. An extreme view, and not one to which we subscribe, might be to define a family as all the tests a researcher might do in a lifetime (see Maxwell & Delaney 1990 and Miller 1981 for discussion), and try to limit the Type I error rate over this family. Controlling error rates over such a family of tests has interesting and humorous implications for biologists' career structures (Morrison 1991). More generally, a family is defined as some collection of simultaneous tests, where a number of hypotheses are tested simultaneously using a single data set from a single experiment or sampling program.

We agree with Hochberg & Tamhane (1987) that unrelated hypotheses (in terms of intended use or content) should be analyzed separately, even if they are not independent of each other. We recommend that each researcher, in a specific analytical situation, must make an *a priori* decision about what a family of tests is; this decision should be based, at least in part, on the relative importance of Type I versus Type II errors.

The other issue is what level to set for family-wise error rate. It is common practice for biologists to set the family-wise Type I error rate to the same level as they use for individual comparisons (e.g. 0.05). This is not easy to justify, especially as it reduces the comparison-wise Type I error rate to very low levels, increasing the probability of Type II errors if any of the H_0s are false. So this is a very conservative strategy and we should consider alternatives. One may be to use a procedure that controls the family-wise error rate but to set a significance level above 0.05. There is nothing sacred about 0.05 (see Section 3.6) and we are talking here about the probability of any Type I error in a collection of tests. Setting this significance level *a priori* to 0.10 or higher is not unreasonable. Another approach is the interesting proposal by Benjamini & Hochberg (1995). They also argued that control of family-wise Type I error rate may be too severe in some circumstances and recommended controlling the false discovery rate (FDR). This is the expected proportion of Type I errors among the rejected hypotheses.

3.4.2 Adjusting significance levels and/or *P* values

Whatever philosophy we decide to use, there will be situations when some control of family-wise Type I error rate will be required. The procedures we will describe here are those which are independent of the test statistic used and are based on adjusting the significance levels for each test downwards to control the family-wise Type I error rate. Note that instead of adjusting significance levels, we could also adjust the *P* values and use the usual significance levels; the two approaches are equivalent.

Bonferroni procedure

This is a general procedure for adjusting significance levels to control Type I error rates in multiple testing situations. Each comparison is tested at

α/c where α is the nominated significance level (e.g. 0.05) and c is the number of comparisons in the family. It provides great control over Type I error but is very conservative when there are lots of comparisons, i.e. each comparison or test will have little power. The big advantage is that it can be applied to any situation where we have a family of tests, so it has broad applicability.

Dunn–Sidak procedure

This is a modification of the Bonferroni procedure that slightly improves power for each comparison, which is tested at $1-(1-\alpha)^{1/c}$.

Sequential Bonferroni (Holm 1979)

This is a major improvement on the Bonferroni procedure where the c test statistics (F, t, etc.) or P values are ranked from largest to smallest and the smallest P value is tested at α/c, the next at $\alpha/(c-1)$, the next at $\alpha/(c-2)$, etc. Testing stops when a non-significant result occurs. This procedure provides more power for individual tests and is recommended for any situation in which the Bonferroni adjustment is applicable.

Hochberg (1988) described a similar procedure that works in reverse. The largest P value is tested at α, rejecting all other tests if this one is significant. If not significant, the next largest is tested against $\alpha/2$, and so on. Shaffer (1995) stated that Hochberg's procedure is slightly more powerful than Holm's.

Resampling-based adjusted *P* values

Westfall & Young (1993a,b) have developed an interesting approach to P value adjustment for multiple testing based around resampling. They defined the adjusted P value as:

$$P_{adj} = P(\min P_{rand} \leq P \mid H_0) \tag{3.10}$$

where P_{rand} is the random P value for any test. Basically, their procedure measures how extreme any particular P value is out of a list of P values from multiple tests, assuming all H_0s are true. Westfall & Young (1993b) argue that their procedure generalizes to Holm's and other methods as special cases and also accounts for correlations among the P values.

3.5 Combining results from statistical tests

We sometimes need to evaluate multiple studies in which statistical analyses have been used to test similar hypotheses about some biological process, such as the effect of a particular experimental treatment. Our interest is in summarizing the size of the treatment effect across studies and also testing an H_0 about whether there is any overall effect of the treatment.

3.5.1 Combining *P* values

Fisher (1954) proposed a method for combining the P values from a number of independent tests of the same hypothesis, even though different statistical procedures, and therefore different H_0s, may have been used (see also Hasselblad 1994, Manly 2001, Sokal & Rohlf 1995). For c independent tests, each producing a P value for the test of a commensurate H_0, the P values can be combined by:

$$-2\sum_{i=1}^{c} \ln(P) \tag{3.11}$$

which is distributed as a χ^2 with $2c$ degrees of freedom. The overall H_0 is that all the H_0s in the collection of tests are true (Sokal & Rohlf 1995). If we reject the overall H_0, we conclude that there is an overall effect of whatever treatment or contrast was commensurate between the analyses. Alternative methods, including ones that weight the outcomes from the different tests differently, are described in Becker (1994) and Manly (2001).

3.5.2 Meta-analysis

The limitation of Fisher's method is that P values are only one piece of information that we use for drawing conclusions from a statistical test. They simply indicate whether we would reject the H_0 at the chosen level of significance. The biological interpretation of that result would depend on the size of the difference or effect, and the sample sizes, so a better approach would incorporate effect sizes, the variances of the effect sizes and sample sizes when combining results from different tests. Such a more sophisticated approach is called meta-analysis. Meta-analysis is used primarily when reviewing the literature on a particular

topic, e.g. competition between organisms (Gurevitch *et al.* 1992), and some overall summary of the conclusions from different studies is required.

Basically, meta-analysis calculates, for each analysis being incorporated, a measure of effect size (Rosenthal 1994, see also Chapters 7 and 8) that incorporates the variance of the effect. These effect sizes from the c different tests are averaged using the sum of the inverse of the variance of each effect size ("inverse variance weighted average": Hasselblad 1994, p. 695). This average effect size can be used as a summary measure of the overall effect of the process being investigated.

Most meta-analyses are based on fixed effects models (see also Chapter 8) where we are assuming that the set of analyses we are combining share some true effect size for the process under investigation (Gurevitch & Hedges 1993). Under this model, the test of H_0 that the true effect size is zero can be tested by constructing confidence intervals (based on the standard normal distribution) for the true average effect size (Gurevitch & Hedges 1993) and seeing if that confidence interval includes zero at the chosen level (e.g. 95%). We can also calculate a measure of homogeneity (Q) for testing whether all c effect sizes are equal. Q is the sum of weighted (by the inverse of the variance of each effect size) squared differences between each effect size and the inverse variance weighted average of the effect sizes. It sounds messy but the computations are quite simple (Gurevitch & Hedges 1993, Hasselblad 1994). Q is distributed as a χ^2 with $c-1$ degrees of freedom. In some cases, the analyses being combined fall into different *a priori* groups (e.g. studies on competition in marine, freshwater and terrestrial environments) and within-group and between-group measures of homogeneity can be calculated (analogous to partitioning the variance in an ANOVA – Chapter 8).

Meta-analysis can be used in any situation where an effect size, and its variance, can be calculated so it is not restricted to continuous variables. Nor is it restricted to fixed effects models, with both random and mixed models possible (Gurevitch & Hedges 1993; see also Chapters 8 and 9). Meta-analyses do depend on the quality of the literature being surveyed. For some studies, not enough information is provided to measure an effect size or its variance. There is also the issue of quality control, ensuring that the design of the studies we have used in a meta-analysis are acceptable, and whether we can combine studies based on experimental manipulations versus those based on weaker survey designs. Nonetheless, meta-analysis is increasing in use in the biological literature and some appreciation of its strengths and weaknesses is important for biologists. One important weakness worth noting is the "file-drawer problem". The database of published papers is highly censored, with non-significant results under-represented, so a meta-analysis of published work should include careful thought about what "population" these published studies represent.

Two detailed texts are Hedges & Olkin (1985) and the volume edited by Cooper & Hedges (1994), although excellent reviews from a biological perspective include Gurevitch & Hedges (1993) and Hasselblad (1994).

3.6 Critique of statistical hypothesis testing

Significance testing, especially null hypothesis significance testing, has been consistently criticized by many statisticians (e.g. Nester 1996, Salsburg 1985) and, in particular, in the recent psychological and educational literature (e.g. Carver 1978, 1993, Cohen 1990, 1994, Shaver 1993, Harlow *et al.* 1997 and chapters therein). Biologists have also questioned the validity of statistical hypothesis testing (e.g. Johnson 1999, Jones & Matloff 1986, Matloff 1991, Stewart-Oaten 1996). A thorough review of this literature is beyond the scope of our book but a brief discussion of these criticisms is warranted.

3.6.1 Dependence on sample size and stopping rules

There is no question that results for classical statistical tests depend on sample size (Chow 1988, Mentis 1988, Thompson 1993), i.e. everything else being the same, larger sample sizes are more likely to produce a statistically significant result and with very large sample sizes, trivial effects

Box 3.5 | Likelihood inference and the likelihood principle

Oakes (1986) described four major schools of statistical inference, three of which we describe in this chapter – Fisherian and Neyman–Pearson hypothesis testing, aspects of both being used by many biologists, and the Bayesian methods based on subjective probabilities. The fourth school is likelihood inference, based on the likelihood function that we outlined in Chapter 2 (see also Royall 1997). There are two important issues involved. First, the evidence that the observed data provide about the hypothesis is represented by the likelihood function, the likelihood of observing our sample data given the hypothesis. Second, the likelihood principle states that two sets of data that produce proportional likelihood functions are equal in terms of evidence about the hypothesis. One of the arguments often used against statistical significance tests is that they violate the likelihood principle.

Likelihood inference is really about relative measures of evidence of support between competing hypotheses so the focus is on the likelihood ratio:

$$\frac{L(\text{data}|H_1)}{L(\text{data}|H_2)}$$

although, as discussed in Chapter 2, we often convert likelihoods to log-likelihoods and the result is a ratio of log-likelihoods. The likelihood ratio can be viewed as a measure of the relative strength of evidence provided by the data in H_1 compared with H_2.

Likelihoods are relevant to both classical and Bayesian inference. Likelihood ratios can often be tested in a classical framework because, under many conditions, the ratio follows a χ^2 distribution. The observed data contribute to a Bayesian analysis solely through the likelihood function and, with a non-informative, uniform prior, the Bayesian posterior probability distribution has an identical shape to the likelihood function.

can produce a significant result. However, while this is true by definition and can cause problems in complex analyses (e.g. factorial ANOVAs) where there are numerous tests based on different df, designing experiments based on *a priori* power considerations is crucial here. Rather than arbitrarily choosing sample sizes, our sample size should be based on that necessary to detect a desired effect if it occurs in the population(s) (Cohen 1988, 1992, Fairweather 1991, Peterman 1990a,b). There is nothing new in this recommendation and we will consider power analysis further in Chapter 7.

The sample size problem relates to the stopping rule, how you decide when to stop an experiment or sampling program. In classical hypothesis testing, how the data were collected influences how we interpret the result of the test, whereas the likelihood principle (Box 3.5) requires the stopping rule to be irrelevant (Oakes 1986). Mayo (1996) and Royall (1997) provide interesting, and contrasting, opinions on the relevance of stopping rules to inference.

3.6.2 Sample space – relevance of data not observed

A well-documented aspect of *P* values as measures of evidence is that they comprise not only the long-run probability of the observed data if H_0 is true but also of data more extreme, i.e. data not observed. The set of possible outcomes of an experiment or sampling exercise, such as the possible values of a random variable like a test statistic, is termed the sample space. The dependence of statistical tests on the sample space violates the likelihood principle (Box 3.5) because the same evidence, measured as likelihoods, can produce different conclusions (Royall 1997). The counter

argument, detailed by Mayo (1996), is that likelihoods do not permit measures of probabilities of error from statistical tests. Measuring these errors in a frequentist sense is crucial to statistical hypothesis testing.

3.6.3 *P* values as measure of evidence

Cohen (1994) and others have argued that what we really want to know from a statistical test is the probability of H_0 being true, given our sample data, i.e. $P(H_0|\text{data})$. In contrast, Mayo (1996) proposed that a frequentist wants to know what is "the probability with which certain outcomes would occur given that a specified experiment is performed" (p. 10). What the classical significance test tells us is the long-run probability of obtaining our sample data, given that H_0 is true, i.e. $P(\text{data}|H_0)$. As Cohen (1994) and others have emphasized, these two probabilities are not interchangeable and Bayesian analyses (Section 3.7), which provide a measure of the $P(H_0|\text{data})$, can produce results very different from the usual significance test, especially when testing two-tailed "point" hypotheses (Berger & Sellke 1987). Indeed, Berger & Sellke (1987) presented evidence that the *P* value can greatly overstate the evidence against the H_0 (see also Anderson 1998 for an ecological example). We will discuss this further in the next section. In reply to Berger & Sellke (1987), Morris (1987) argued that differences between *P* values and Bayesian posteriors will mainly occur when the power of the test is weak at small sample sizes; otherwise *P* values work well as evidence against the H_0. Reconciling Bayesian measures and *P* values as evidence against the H_0 is still an issue of debate among statisticians.

3.6.4 Null hypothesis always false

Cohen (1990) and others have also argued that testing an H_0 is trivial because the H_0 is always false: two population means will never be *exactly* the same, a population parameter will never be *exactly* zero. In contrast, Frick (1995) has pointed out an H_0 can be logically true and illustrated this with an ESP experiment. The H_0 was that a person in one room could not influence the thoughts of a person in another room. Nonetheless, the argument is that testing H_0s is pointless because most common H_0s in biology, and other sciences, are always false. Like Chow (1988, 1991) and Mulaik *et al.* (1997), we argue that the H_0 is simply the complement of the research hypothesis about which we are trying to make a decision. The H_0 represents the default (or null) framework that "nothing is happening" or that "there is no effect" (3.1.1). A rejection of the H_0 is not important because we thought the H_0 might actually be true. It is important because it indicates that we have detected an effect worth reporting and investigating further. We also emphasise that H_0s do not have to be of the "no effect" form. There may be good reasons to test H_0s that a parameter equals a non-zero value. For example, in an environmental monitoring situation, we might compare control and impact locations to each other, and look for changes through time in this control–impact difference. We might find that two locations are quite different from each other as a result of natural processes, but hypothesize that a human activity will change that relationship.

3.6.5 Arbitrary significance levels

One long-standing criticism has been the arbitrary use of $\alpha = 0.05$ as the criterion for rejecting or not rejecting H_0. Fisher originally suggested 0.05 but later argued against using a single significance level for every statistical decision-making process. The Neyman–Pearson approach also does not rely on a single significance level (α), just a value chosen *a priori*. There is no reason why all tests have to be done with a significance level fixed at 0.05. For example, Day & Quinn (1989) have argued that there is nothing sacred about 0.05 in the context of multiple comparisons. Mapstone (1995) has also provided a decision-making framework by which the probabilities of Type I and Type II errors are set based on our assessment of the cost of making the two types of error (Section 3.2.2). The point is that problems with the arbitrary use of 0.05 as a significance level are not themselves a reason to dismiss statistical hypothesis testing. Irrespective of which philosophy we use for making statistical decisions, some criterion must be used.

3.6.6 Alternatives to statistical hypothesis testing

In the discussions on significance testing, particularly in the psychological literature, three general

alternatives have been proposed. First, Cohen (1990, 1994) and Oakes (1986) and others have argued that interval estimation and determination of effect sizes (with confidence intervals) is a better alternative to testing null hypotheses. While we encourage the use and presentation of effect sizes, we do not see them as an alternative to significance testing; rather, they are complementary. Interpreting significance tests should always be done in conjunction with a measure of effect size (e.g. difference between means) and some form of confidence interval. However, effect sizes by themselves do not provide a sensible philosophical basis for making decisions about scientific hypotheses.

Second, Royall (1997) summarized the view that likelihoods provide all the evidence we need when evaluating alternative hypotheses based on the observed data. Finally, the Bayesian approach of combining prior probability with the likelihood function to produce a posterior probability distribution for a parameter or hypothesis will be considered in the next section.

In summary, biologists should be aware of the limitations and flaws in statistical testing of null hypotheses but should also consider the philosophical rationale for any alternative scheme. Does it provide us with an objective and consistent methodology for making decisions about hypotheses? We agree with Dennis (1996), Levin (1998), Mulaik *et al.* (1997) and others that misuse of statistical hypothesis testing does not imply that the process is flawed. When used cautiously, linked to appropriate hypotheses, and combined with other forms of interpretation (including effect sizes and confidence intervals), it can provide a sensible and intelligent means of evaluating biological hypotheses. We emphasize that statistical significance does not necessarily imply biological importance (Box 3.4); only by planning studies and experiments so they have a reasonable power to detect an effect of biological importance can we relate statistical and biological significance.

3.7 Bayesian hypothesis testing

One approach that may provide a realistic alternative to classical statistical hypothesis testing in some circumstances is Bayesian methodology. As we discussed in Chapter 2, the Bayesian approach views population parameters (e.g. means, regression coefficients) as random, or at least unknown, variables. Bayesians construct posterior probability distributions for a parameter and use these probability distributions to calculate confidence intervals. They also use prior information to modify the probability distributions of the parameters and this prior information may include subjective assessment of prior probabilities that a parameter may take specific values.

The Bayesian approach rarely incorporates hypothesis testing in the sense that we have been discussing in this chapter and Bayesian do not usually evaluate alternative hypotheses or models with a reject/accept decision framework. They simply attach greater or lesser favor to the alternatives based on the shape of the posterior distributions. Nonetheless, there are some formal ways of assessing competing hypotheses using Bayesian methods.

We might, for example, have two or more rival hypotheses (H_1, H_2,…H_i); in the classical hypothesis testing framework, these would be H_0 and H_A, although a null hypothesis of no effect would seldom interest Bayesians. We can then use a similar version of Bayes theorem as described for estimation in Chapter 2:

$$P(H_1 \mid \text{data}) = \frac{P(\text{data} \mid H_1) P(H_1)}{P(\text{data})} \tag{3.11}$$

where $P(H_1 \mid \text{data})$ is the posterior probability of H_1, $P(H_1)$ is the prior probability of H_1 and $P(\text{data} \mid H_1)/P(\text{data})$ is the standardized likelihood function for H_1, the likelihood of the data given the hypothesis. For example, we could test an H_0 using the Bayesian approach by:

$$\text{posterior probability of } H_0 = \text{likelihood of data given } H_0 \cdot \text{prior probability of } H_0 \tag{3.12}$$

The posterior probability is obtained by integrating (if the parameter in the H_0 is continuous) or summing (if discrete) under the posterior probability distribution for the range of values of the parameter specified in the H_0. For continuous parameters, the procedure is straightforward for directional (composite) hypotheses, e.g. H_0: θ less than some specified value, but difficult for a point

(simple) hypothesis, e.g. H_0: θ equals some specified value, because we cannot determine the probability of a single value in a probability distribution of a continuous variable.

We can present the relative evidence for H_0 and H_A as a posterior odds ratio:

$$\frac{P(H_0|\text{data})}{P(H_A|\text{data})} \tag{3.13}$$

i.e. the ratio of the posterior probabilities, given the data, of the competing hypotheses (Reckhow 1990). This posterior odds ratio is also the product of the prior odds ratio with a term called the Bayes factor (Barnett 1999, Ellison 1996, Kass & Raftery 1995, Reckhow 1990). If the two hypotheses were considered equally likely beforehand, then the Bayes factor equals the posterior odds ratio. If the prior odds were different, then the Bayes factor will differ from the posterior odds ratio, although it seems that the Bayes factor is primarily used in the situation of equal priors (Kass & Raftery 1995). Both the Bayes factor and the posterior odds ratio measure the weight of evidence against H_A in favor of H_0, although the calculations can be reversed to measure the evidence against H_0.

When both hypotheses are simple (i.e. θ equals a specified value), the Bayes factor is just the likelihood ratio (Box 3.5):

$$B = \frac{L(\text{data}|H_0)}{L(\text{data}|H_A)} \tag{3.14}$$

where the numerator and denominator are the maxima of the likelihood functions for the values of the parameter specified in the hypotheses. When one or both hypotheses are more complex, the Bayes factor is still a likelihood ratio but the numerator and denominator of Equation 3.14 are determined by integrating under the likelihood functions for the range of parameter values specific in each hypothesis (Kass & Raftery 1995). We are now treating the likelihood functions more like probability distributions. For complex hypotheses with multiple parameters, this integration may not be straightforward and the Monte Carlo posterior sampling methods mentioned in Chapter 2 might be required.

To choose between hypotheses, we can either set up a decision framework with an *a priori* critical value for the odds ratio (Winkler 1993) or, more commonly, use the magnitude of the Bayes factor as evidence in favor of a hypothesis. A simpler alternative to the Bayes factor is the Schwarz criterion (or Bayes Information Criterion, BIC), which approximates the log of the Bayes factor and is easy to calculate. Ellison (1996) has provided a table relating different sizes of Bayes factors (both as $\log_{10}B$ and $2\log_e B$) to conclusions against the hypothesis in the denominator of Equation 3.14. Odds and likelihood ratios will be considered in more detail in Chapters 13 and 14.

Computational formulae for various types of analyses, including ANOVA and regression linear models, can be found in Box & Tiao (1973), while Berry & Stangl (1996) have summarized other types of analyses. Hilborn & Mangel (1997) focused on assessing the fit of models to data using Bayesian methods. In a fisheries example, they compared the fit of two models of the dynamics of hake off the coast of Namibia where one model was given a higher prior probability of being correct than the second model. As another example, Stow *et al*. (1995) used Bayesian analysis to estimate the degree of resource dependence (ϕ) in lake mesocosms with different ratios of grazing *Daphnia*. Using a non-informative prior, a high value of ϕ, indicating much interference among the predators, had the highest posterior probability. Stow *et al*. (1995) pointed out that, in contrast, classical statistical analysis would only have shown that ϕ was significantly different to some hypothesized value. A third example is Crome *et al.* (1996), who compared Bayesian (with a range of prior distributions) and classical linear model analyses of a BACI (Before-After-Control-Impact) design assessing the effects of logging on birds and mammals in a north Queensland rainforest. Although the two approaches produced similar conclusions for some variables, the posterior distributions for some variables clearly favored some effect sizes over others, providing more information than could be obtained from the classical test of a null hypothesis.

When classical P values [$P(\text{data}|H_0)$] are compared to Bayes factors or Bayesian posterior probabilities [$P(H_0|\text{data})$], the differences can be marked, even when H_0 and H_A are assigned equal prior probabilities (i.e. considered equally likely).

Berger & Sellke (1987) and Reckhow (1990) argued that the differences are due to the *P* value being "conditioned" on the sample space, including an area of a probability distribution that includes hypothetical samples more extreme than the one observed (Section 3.6.2). In contrast, the Bayesian posterior probability is conditioned only on the observed data through the likelihood. The differences between *P* values and Bayesian posterior probabilities seem more severe for two-tailed testing problems (Casella & Berger 1987), where the *P* value generally overstates the evidence against H_0, i.e. it rejects H_0 when the posterior probability suggests that the evidence against H_0 is relatively weak. Nonetheless, *P* values will mostly have a monotonic relationship with posterior probabilities of H_0, i.e. smaller *P* values imply smaller posterior probabilities, and for one-tailed tests (e.g. ANOVA *F*-ratio tests), there may be equivalence between the *P* values and posterior probabilities for reasonable sorts of prior distributions (Casella & Berger 1987). So it may be that the relative sizes of *P* values can be used as a measure of relative strength of evidence against H_0, in the sense that they are related to Bayesian posterior probabilities (but see Schervish 1996; also Royall 1997 for alternative view).

One of the main difficulties classical frequentist statisticians have with Bayesian analyses is the nature of the prior information (i.e. the prior probabilities). We discussed this in Chapter 2 and those issues, particularly incorporating subjective probability assessments, apply just as crucially for Bayesian hypothesis testing.

So, when should we adopt the Bayesian approach? We have not adopted the Bayesian philosophy for the statistical analyses described in this book for a number of reasons, both theoretical and practical. First, determining prior probabilities is not straightforward in those areas of biology, such as ecology, where much of the research is still exploratory and what happened at other times and places does not necessarily apply in a new setting. We agree with Edwards (1996) that initial analyses of data should be "journalistic", i.e. should not be influenced by our opinions of what the outcome might be (prior probabilities) and that there is an argument that using prior (personal) beliefs in analyses should not be classified as science. While Carpenter (1990) and others have argued that the prior probabilities have relatively little influence on the outcome compared to the data, this is not always the case (Edwards 1996). For the types of analyses we will discuss in this book, any prior information has probably already been incorporated in the design components of the experiment. Morris (1987) has argued that *P* values are interpretable in well-designed experiments (and observational studies) where the power to detect a reasonable H_A (effect) has been explicitly considered in the design process. Such a well-designed experiment explicitly considering and minimizing Type I and Type II errors is what Mayo (1996) would describe as a severe test of an hypothesis. Second, treating a population parameter as a random variable does not always seem sensible. In ecology, we are often estimating parameters of real populations (e.g. the density of animals in an area) and the mean of that population is a fixed, although unknown, value. Third, Bayesian analyses seem better suited to estimation rather than hypothesis testing (see also Dennis 1996). Some well-known Bayesian texts (e.g. Box & Tiao 1973, Gelman *et al.* 1995) do not even discuss hypothesis testing in their Bayesian framework. In contrast, the philosophical position we take in this book is clear. Advances in biology will be greatest when unambiguously stated hypotheses are tested with well-designed sampling or preferably experimental methods. Finally, the practical application of Bayesian analyses is not straightforward for complex analyses and there is little software currently available (but see Berry 1996, Berry & Stangl 1996 and references in Ellison 1996). We suspect that if biologists have enough trouble understanding classical statistical analyses, Bayesian analyses, with their reliance on defining probability distributions and likelihood functions explicitly, are more likely to be misused.

There are some circumstances where the Bayesian approach will be more relevant. In environmental management, managers often wish to know the probability of a policy having a certain outcome or the probabilities of different policies being successful. Whether policies are significantly different from one another (or different from some hypothesized value) is not necessarily

helpful and Bayesian calculation of posterior probabilities of competing models might be appropriate. Hilborn & Mangel (1997) also emphasize Bayesian methods for distinguishing between competing models. This in itself has difficulties. Dennis (1996) correctly pointed out the danger of various interest groups having input into the development of prior probabilities, although we have argued earlier (Section 3.2.2) that such negotiation in terms of error rates in the classical decision-making framework should be encouraged. One-off, unreplicated, experiments might also be more suited to Bayesian analyses (Carpenter 1990) because the long-run frequency interpretation doesn't have much meaning and the probability of a single event is of interest.

Bayesian approaches are being increasingly used for analyzing biological data and it is important for biologists to be familiar with the methods. However, rather than simply being an alternative analysis for a given situation, the Bayesian approach represents a different philosophy for interpreting probabilities and we, like Dennis (1996), emphasize that this must be borne in mind before it is adopted for routine use by biologists.